EFFICACY IN AUTOMATED LANGUAGE TRANSLATORS

MAJ Ian McCulloh,* 2LT Jillian Morton, 2LT Jennifer K. Jantzi, 2LT Amy M. Rodriguez, and
LTC John Graham
United States Military Academy
West Point, New York 10996

ABSTRACT

This paper suggests an improved measure for evaluating the usefulness of automated machine language translators. With the Global War on Terror (GWOT), the Army has increasing interest and need for accurate language translation more than ever. Today, there are approximately 20,000 linguists with language training in either the Active Duty or Reserve components of the U.S. Army. Coalition operations and U.S. presence in Iraq, Kuwait, and other areas in the Middle East require Arabic translation. Unfortunately, the Army has never been able to maintain the number of linguists it needs, particularly in the hard-to-fill, low-density languages (Dunn 1).

Previous evaluations of machine translations usually rely on word error rate. Word error rate is calculated by adding the number of insertions, deletions, or substitutions of words in one language to another language (LaRocca 1). The problem with this evaluation method is that it does not take into account human cognition or context. In other words, a machine translation might have a high word error rate but the user can still understand the "gist." In short, past methods of evaluation do not consider user knowledge or experience. Machine translation systems should be rated not in terms of their word error rate but in terms of human comprehension and usefulness, which is some function of word translation, syntax translation, and semantic interpretation.

This study introduces a new method of evaluating human comprehension in the context of machine translation using a language translation program known as the Forward Area Language Converter (FALCon). A study was conducted where participants received seven translated articles in a random order. For each of the seven articles, the participants received a set of corresponding comprehension questions. The goal of the questions was to gear the reader toward intelligence gathering and to see if he could grasp main concepts and details.

The results of this study suggest that word error rate is not an effective measure of the usefulness of a machine language translator. Comprehension tests perform better at evaluating a human's understanding of a translated document. This study further indicates strengths and weaknesses in each translator.

BACKGROUND

Of the 20,000 linguists serving in either the Active Duty or Reserve components of the U.S. Army, less than half belong to the Military Intelligence branch (LaRocca 1). More than ever, the Army has increasing interest and need for accurate language translation especially with the Global War on Terror (GWOT). Unfortunately, the Army is not able to maintain the number of linguists it requires.

The U.S. Army operating on foreign soil in both peacekeeping and combat operations cannot afford to ignore the language barrier. In addition to communicating with the populace or gleaning intelligence from enemy documents, the Army is increasingly cooperating with coalition forces, which also introduces a variety of languages and thus other language barriers. To overcome this problem, the most obvious solution would seem to hire more translators. However, there are disadvantages to this. For one, the quality of translations can be mixed, some translators may be better communicators than others. More importantly, hiring translators can be downright dangerous. Translators could be operating in conjunction with the enemy or be providing false information to that effect. To alleviate the burden of language translation, many are looking toward machine translation, as a means to augment linguists in theater.

Opponents of automated machine translation cite the multiple errors that occur and thus conclude that machine translation does not add significant benefits. Previous evaluations of machine translations usually rely on word error rate. Word error rate, designed to measure accuracy, is calculated by adding the number of insertions, deletions, or substitutions of words in one language to another language (LaRocca 1). Usually word error rate is determined using a computer program which calculates using the following equation, $(n-e)/n$, where n is the number of characters in the ground-truth file, and e is every character inserted, substituted, or deleted.

Unfortunately, this evaluation method does not take into account human cognition or context. A machine translation might have a high word error rate, but the user can still understand the "gist." The current standard for evaluation does not consider user knowledge or experience. Machine translation systems should not be rated in terms of their word error rate, but rather in terms

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 NOV 2006	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Efficacy In Automated Language Translators	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **United States Military Academy West Point, New York 10996**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002075.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**4**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

of human comprehension and usefulness. Usefulness is a function of the human interpretation of the machine translated text. This usefulness function is referred to as the "gist" of the text.

This study introduces a new method of evaluating human comprehension in the context of machine translation. The Forward Area Language Converter (FALCon) was used as the machine language translator for the study. The FALCon works by converting documents into digital images via scanner, and then converting those images to electronic text by use of the Optical Character Recognition, (OCR). Foreign text is then converted to English using the Machine Translator. In all, the FALCon can negotiate 61 languages though some languages do not have OCR capacity. Foreign and English texts are searched for keywords from a list that the operator selects based on need or context.

The FALCon has two translation systems; Cybertran and Transphere. Cybertran negotiates text at the literal level while the Transphere attempts to incorporate syntactical meaning in its translations. Syntax refers to the rules of sentence formation; the component of the mental grammar that represents speakers' knowledge of the structure of phrases and sentences. For example, in Spanish, syntax includes a noun followed by an adjective: i.e. "Tengo la camisa negra" which taken literally in English means: "I have the shirt black" Cybertran would translate in this manner. However, Transphere would syntactically adjust that same sentence to: "I have the black shirt." Cybertran leaves the reconstruction of a sentence and the context to the analyst. This relies on the analysts' learned understanding of the foreign language sentence form. Transphere, on the other hand, attempts to incorporate syntactical rules into the translation. In this way, Transphere attempts to make the sentence structure more similar to the English language, however, it introduces noise into the process.

In addition to structure of language, readers also rely on schema to increase the understanding of text. Schemas help linguists understand the story structure (Braintree). Though literal translation is a priority for the reader, the coherent meaning constructed by the reader will often reflect a reader's prior experience. The recall protocols of foreign language students demonstrate that though students can often recognize words, they seriously misread or misconstrue their meaning within different contexts (Swaffar 123). The more familiar a linguist is with the structure of a language, the better they will be at grasping the "gist."

Others argue that machine translators have a poor reputation because people have the wrong expectations of what machine translators are capable of. They should be seen as a tool that can be used to assist in translating because "even if machine translation systems can never duplicate human translations, can't they at least generate output that is understandable and useful (Myers 2)?"

It is hypothesized that semantic machine translations (Transphere) will result in better reading comprehension as the reader begins to develop an implicit understanding of the sentence structure. A second hypothesis proposes that over time and with practice, the literal machine translation system (Cybertran) will produce a reading comprehension curve that increases over time while the semantic translation system will initially be higher due to its resemblance of the English language, but then over time, will level off because of the noise it introduces. This concept is illustrated in Figure 1.

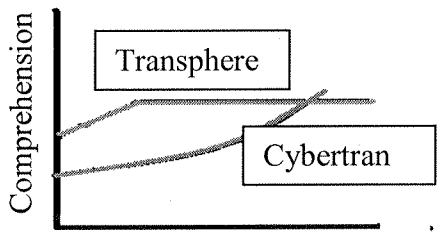

Figure 1. Comprehension over Time.

METHODOLOGY

An experiment was conducted to compare human comprehension using the two different machine translation systems. The participants for this experiment included 48 freshmen from the United States Military Academy enrolled in the General Psychology course, PL100. The investigators used the CyberTran and Transphere programs to translate seven Arabic electronic documents/websites. Each of the seven articles also had an English translation prepared by a human linguist. A series of corresponding comprehension questions were developed to evaluate a subject's comprehension of an article. These comprehension questions were developed in conjunction with the Department of Foreign Languages at the U.S. Military Academy. They were specifically designed to assess comprehension in the same manner as a language instructor would assess comprehension in an intermediate language course at the Military Academy.

Several actions were taken to reduce variability in the experiment. The experiment was a between subjects design. The participants were equally divided into two groups. The FALCon program used electronic texts to eliminate potential error introduced with the OCR. The Arabic documents were the same for each experimental group, except for the type of translation used to convert them to English. One group received the seven Arabic documents translated into English using the Transphere

system. The second group received the documents translated into English using the Cybertran system. All subjects received the articles in a random order.

For each of the seven articles, the subjects received a set of corresponding comprehension questions. The questions were the same for each participant, despite the condition. The participants were instructed to read and answer the comprehension questions to the best of their ability. Once they finished answering the questions, the participants were given the master English copy of the article so that they could compare this document to the translation produced by either the Transpere or Cybertran. The intent was to determine if they could better understand syntax and vocabulary over time.

Each test for the seven articles was designed in the same format consisting of two multiple choice questions, one fill-in the blank question, one true/false question and a two-part question wherein subjects were required to re-structure a translated sentence from both the Transphere and Cybertran translators. The goal of putting the questions in a particular order was to gear the reader toward intelligence gathering and to see if he could grasp main concepts and details, and overtime (though not yet evaluated in this study), have him answer these kind of questions without being prompted. The first question is always a main idea question, to gauge the reader's overall understanding and force him to think about the main concepts of the subject before he answers smaller questions. The second question was a detail within the article that was important to the overall article. This detail either asked for a key person or place within the article. The fill-in the blank question was another detail, but not as specific as the multiple choice question; it would ask for how often an occurrence happened or who a significant person was (based on position more than specific name). The true/false question was geared to be a little tricky to readers to see if they truly understood a broad concept of the article. The question would entail a detail that encompassed the overall significance of the article. For instance, one question read, "True or False: Each of these Iraqis think that the establishment of a government will solve the problems in the country." Listening to current press reports in general, most people would choose false, but those who read and understood this particular article would correctly answer true to this question, thus the question aids (but does not define) the assessment of one's understanding of the article. The last two questions are meant to gauge which article would be more conducive to translation from "translator garb" to understandable English. This is done by asking the subjects to re-construct two sentences, one from each of the translators, into a coherent sentence. This last question attempts to evaluate whether human understanding and interpretation can fill in the gaps of a poorly-written document by reconfiguring the sentence in his own words while retaining the original meaning of the article. Each test received a score based on a 24 point scale, much like a teacher would grade a test for students. Each multiple choice and true/false question was worth 3 points, while each fill-in the blank and short answer question was worth 5 points. Partial credit was awarded to those answers that showed some valid comprehension of the material.

ANALYSIS AND RESULTS

There was insufficient power to show statistically significant learning for overall test scores. If learning were present, the results would show an increase of correct responses over time. However, it appears that for certain questions, correct responses may increase over time. The questions that showed an increase in correct responses over time were the ones which required specific answers, such as the multiple choice for detail, fill-in the blank, and true/false. Figure 2 shows the average scores for participants over the span of the test from the first article they were given to the last article for the specific-response type questions.

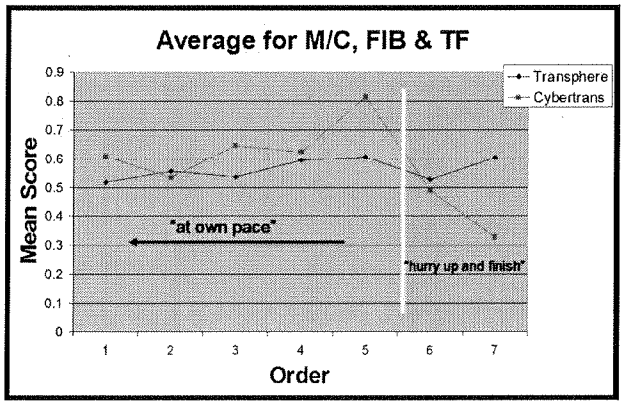

Figure 2. Average Test Score for Multiple Choice, Fill-in-the-Blank, and True/False Questions.

The vertical line in Figure 2 separates the final two tests, wherein subjects had to rush to finish the test within the given time period. Until that time, the number of correct Cybertran responses were improving consistently, while the Transphere scores remained steady.

Cybertran's word-for-word translations seemed to make picking out details within the article a simpler task for subjects than the Transphere's translations. Showing even more evidence of the difference between the two types of translation is the graph of the multiple choice detail questions over time, as seen in Figure 3. While the two types of translations started at the same place for the first question, each subsequent question showed a steady rise in percent answered correctly for subjects reading the

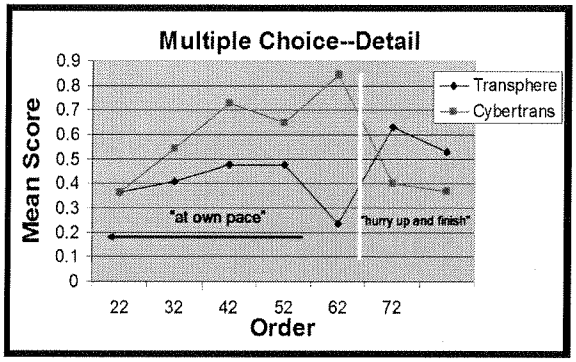

Figure 3. Average Test Score for Multiple Choice Detail Questions.

Cybertran translations. However, it appears that when subjects answer in haste, those reading Cybertran translations struggled. This may be a result of their inability to look back in the article for key words. The Transphere most likely did better, because it gives readers a better idea of the broad sense of the article, so they could still venture a good guess even when rushed.

The difference in overall scores for articles and for individual questions was also analyzed. Figure 4 illustrates the difference in correct responses between each translator by questions asked. The Transphere

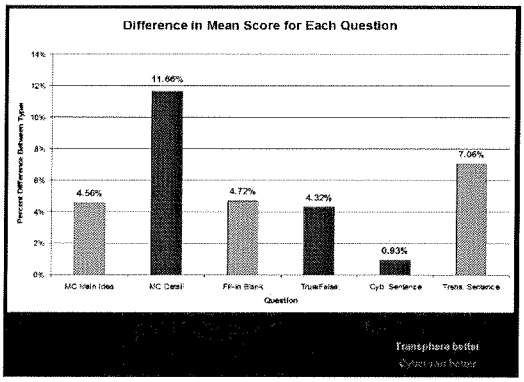

Figure 4. Difference in Mean Score for Each Question.

System performed better on broad idea questions, which means that instead of asking for a particular person or fact, they ask for an idea or underlying concept.

CONCLUSION AND FUTURE STUDY

The results of this study have brought a few key points for consideration in machine translation. First, human understanding is not a factor to be ignored in gauging the usefulness of such translators. Secondly, the type of translation used can depend on the type of information needed, whether it is key people and places or the general plans or opinions. An even better method would combine the two types (probably through human interpretation) to have one complete translation with both key details and the right concepts. Combining the strengths of the two types is especially important in developing a training strategy to employ translators like the FALCon for intelligence gathering. If a soldier can be trained to interpret machine translations within a relatively short period of time, then the lengthy process of finding a linguist and sending and receiving a document can be eliminated, and articles can be processed and interpreted in a very short time by a member of the unit.

The learning factor requires further evaluation. Can people be trained to understand the machine translation better? If so, is one of the two types of translators easier to learn. To evaluate these questions, subjects could perform numerous test sessions over the period of a few months instead of working for only an hour. During this time, subjects may slowly adapt to a different kind of test that would change from some multiple choice questions at the beginning to short answer and eventually to straight essay at the end, wherein they would attempt to touch on all the same key concepts from the first tests. If a person could obtain all the important information without being guided by questions, then that would truly test his understanding of the article and prove the translator's value to the intelligence community. To further asses the learning involved, the subject could be made aware of the exact rules that go into each translation system and then be given the tests.

REFERENCES

Army Research Laboratory. (2004) "Training: Forward Area Language Converter (FALCon)."

Dunn, Kenneth. "Language tools- automated translation." Military Intelligence Professional Bulletin (Jan-March, 2003). [website online]. Available from: http://www.findarticles.com/p/articles/mi_m0IBS/is_1_29/ai_97822089; Accessed 15 Sept. 2005,

http://www.brainconnection.com/content/5_1. Paragraph Comprehension: The Connection to Reading Skills (May 2001).

Human Intelligence and Counterintelligence (CI) Support Tools (HICIST). Available from [website online] Available from: http://www.globalsecurity.org/intell/systems/hicist.htm; Accessed 15 2005.

Myers, Stephen. "Can Computers Translate?" Computing Japan Magazine

Swam, Kathleen. (1999). "FALCon: Evaluation of OCR and Machine Translation Paradigms." US Army Research Laboratory.

Tanner, Simon. (2004) "Deciding Whether Optical Character Recognition is Feasible." King's Digital Consultancy Services.